

# Use Cases for CDF dCache Analysis Diskpool

Ray Culbertson  
FNAL

# Unofficial Levels of DH Organization

---

## Production

- ♦ collaboration-wide, raw and reconstructed data
- ♦ not addressed here

## Physics Groups

- ♦ 5 groups: B, Top, Exotic, QCD, Ewk
- ♦ data ntuple making
  - > Stntuple, B-Stntuple and TopNtuple
  - > groups have priorities, special needs
- ♦ large MC datasets, and their ntuples

## Small Analysis Groups

- ♦ data ntuple sub-skims or personal, derived ntuples
- ♦ analysis specific MC
- ♦ special study samples

*Diskpool will be used by Physics and Analysis groups*

# Ntuple Use

---

## Producing Ntuples

- ♦ groups of users submit organized jobs to CAF farms
- ♦ ntuple datasets are medium (500GB) to large (2TB)
- ♦ totals for  $1\text{fb} = 10^3 \text{ of TB}$
- ♦ output checked for completeness, cataloged, written to diskpool

## Validation, Aging, Recovery, Concatenation

- ♦ official, but limited detail, validation
- ♦ validation through user experience
  - > ex: need more muon variables...
  - ♦ recovering files
    - > ex: still working on recovering certain read errors
- ♦ concatenation: takes time, needs a large staging area
  - ♦ first, fast look

## Backup

- ♦ once stable (1-3 months) primary data ntuples are sent to tape

*Diskpool can be used to stage and stabilize ntuple datasets*

# Analysis Group Use

---

## Organization

- ♦ Offline gives Physics groups space as needed
- ♦ Physics groups gives analysis groups space as needed

## Use

- ♦ smaller signal MC's
- ♦ skimmed ntuples
- ♦ study samples
- ♦ lifetime of these areas is 3 to 12 months
- ♦ little of this is backed up - single use, limited lifetime

*Diskpool can be used by Physics/Analysis groups as  
a large, flexible scratch area*

# User Notes

---

Some observations from a non-expert...

## Flexibility

- Offline was able to add disk quickly on short notice
  - > even on Thanksgiving eve...
  - additions occurred while we were writing full speed

## Uniformity of one effective partition

- allows trivial scripting for writing and reading
- allows trivial monitoring of space, and who is using what
- easy to pack data densely
- easy to shuffle data over multi-TB space

## Thoughts

- known quantity - already reading ntuples from general dCache
  - OTOH, I do not know when/how large-scale ntuple reading will fail
  - dCache has glitched in the past - needs to be reliable (downtimes too)
  - reading via SAM may alter current procedures and observations
- Diskpool is a very convenient, familiar way to provide large disk space to Physics/Analysis groups*